

《心理学报》审稿意见与作者回应

题目：两种新的多维计算机化分类测验终止规则

作者：任赫，陈平

第一轮

审稿人 1 意见：

这篇文章，基于多维的项目反应理论和计算机分类测验，提出了两种新的多维计算机分类测验终止规则，分别是基于马氏距离的多维序贯似然比规则和随机缩减的多维广义似然比规则，同时介绍了四种已有的多维计算机分类测验并进行了相应的模拟研究。文章选题新颖，但方法的理论缺乏严谨性，具体问题如下。

回应：非常感谢您对本研究及计算机化分类测验终止规则领域给予的正面评价。关于“方法的理论缺乏严谨性”的问题，我们将在下文中就具体建议给予回应。

意见 1：文中第七页，在 M-GLR 方法介绍中，没有对 θ_1 , θ_2 进行定义，含义不清楚；请确定公式(13)是否准确，如此定义与后文的判断准则是矛盾的；此外，判断准则描述的不清晰。

回应：十分感谢您的建议。公式（13）中包含 $\theta_1 \in \Theta_m$ 及 $\theta_2 \in \Theta_n$ 的描述，表示 θ_1 为“达标”的被试的能力空间 Θ_m 中的任一值， θ_2 为“未达标”的被试的能力空间 Θ_n 中的任一值。我们已将上述内容补充到公式（13）下面，详见修改稿的第 8 页。

对于您所提到的“与后文的判断准则是矛盾的”，如果您这里的“矛盾”指的是后文的 M-SCGLR 中出现了 $\hat{\theta}_u$ 和 $\hat{\theta}_l$ （即您在意见 2 中所指出的问题），很抱歉在 M-SCGLR 的描述中确实存在错误，我们已经进行了相应修改。详见对您意见 2 的回应及修改稿的第 10 页。

在判断准则的描述方面，尽管 M-GLR、C-SPRT 和 P-SPRT 规则在统计量 C_{ij} 的构造上有所区别，但是这些规则在判断准则上是一致的。因此，我们没有在 M-GLR 方法的介绍中再次呈现判断准则（即公式（11））。为帮助您理解，我们制作了下述关于判断准则的流程图。此外，我们对公式（11）进行了调整，详见修改稿的第 7 页。

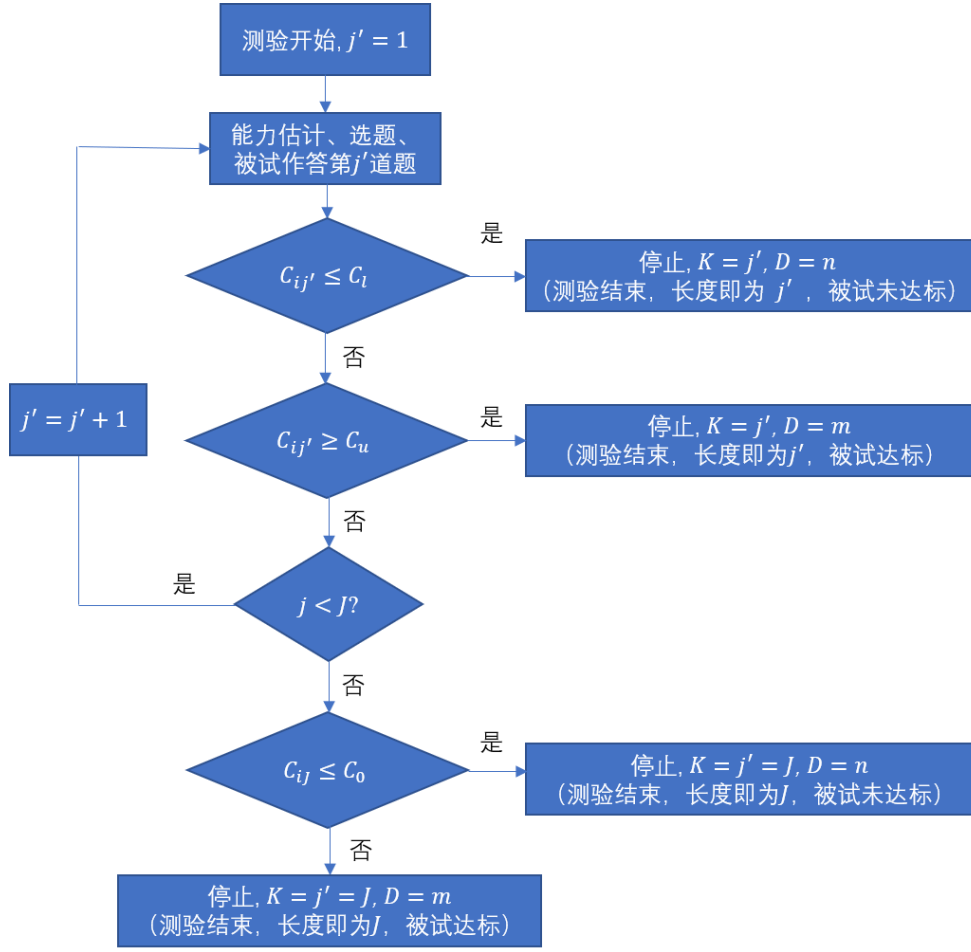


图 1 C-SPRT、P-SPRT 及 M-GLR 的判断准则的流程图

意见 2：文中第十页，**M-SCGLR** 应该是基于前文所述的 **M-GLR**，在前面的描述中说不需要确定 $\hat{\theta}_u$ 和 $\hat{\theta}_l$ ，那么公式(22)和公式(23)中的 $\hat{\theta}_u$ 和 $\hat{\theta}_l$ 又是如何得出的，这是否正确？如果有误，请进行修改。

回应：十分抱歉这里确实存在笔误。在公式（22）及（23）中确实不应出现 $\hat{\theta}_u$ 和 $\hat{\theta}_l$ ，原文中呈现的两个公式表示的是 **M-SCSPRT** 方法中对期望和方差的计算方式，而不是 **M-SCGLR** 方法的。更正后的公式应为，

$$\mathbb{E}_{\theta}(C_{ij}|C_{ij'}) = C_{ij'} + \sum_{j'+1}^J \mathbb{E}_{\theta} \left(\log \frac{\sup_{\theta_1 \in \Theta_m} [L(\theta_1 | \mathbf{y}_{ij'})]}{\sup_{\theta_2 \in \Theta_n} [L(\theta_2 | \mathbf{y}_{ij'})]} \right), \quad (22)$$

$$\text{Var}_{\theta}(C_{ij}|C_{ij'}) = \sum_{j'+1}^J \text{Var}_{\theta} \left(\log \frac{\sup_{\theta_1 \in \Theta_m} [L(\theta_1 | \mathbf{y}_{ij'})]}{\sup_{\theta_2 \in \Theta_n} [L(\theta_2 | \mathbf{y}_{ij'})]} \right). \quad (23)$$

我们已将上述修改更新到公式（22）与（23）中，再次对出现这样的低级错误感到抱歉。

详见修改稿的第 10 页。

意见 3：文中第九页，2.3.1 中“记被试*i*作答完第*j*道题目后得到的能力估计值为 $\hat{\theta}_{ij}$ ，那么他/她在做作答完*j'*道题目后可得到一组能力估计值 $\hat{\theta}_{ij}(j = 1, 2, \dots, j')$ ”，此处如此计算的协方差 $\Sigma_{j'}$ ，并不是不同维度能力之间的相关系数。该相关矩阵只是一个被试作答不同题目后能力估计值的相关性。不同能力之间的相关性应是被试总体层面的，被试个体各个能力不存在关系。

回应：我们非常认同您的观点。很抱歉原文中对 Mahalanobis-SPRT 方法的介绍存在一定错误。 $\Sigma_{j'}$ 的确为一个被试作答*j'*道题目后，得到的*j'*个能力估计值的协方差矩阵。我们对原文中 Mahalanobis-SPRT 方法的介绍及其优缺点分析进行了整体的修改，详见修改稿的第 9 页。

意见 4：文中第十二页，((3)终止规则与分界曲线)，设置分界边界分类曲线 $g(\theta)$ 时， $g(\theta)$ 的选取会对结果产生影响，能否考虑分类曲线的不同设置方法并进行比较。

回应：非常感谢您的建议。根据 Nydick (2013)¹的研究，能力分界函数 $g(\theta)$ 主要可以分为补偿性函数及非补偿性函数两类。举例来说，

$$g(\theta) = \theta_1 + \theta_2,$$

即为一个补偿性的分界函数。相对地，

$$g(\theta) = \begin{cases} \theta_1, & \text{若 } \theta_2 \geq 0 \\ \theta_2, & \text{若 } \theta_1 > 0 \end{cases}$$

就是一个非补偿性的分界函数。

为考虑不同分界函数对结果的影响，我们对实验设计部分进行修改并重新进行了模拟研究。在保留原有的补偿性分界函数 $g(\theta) = \theta_1 + \theta_2$ 的基础上，我们增加了上文所述的这一非补偿性的分界函数。详见修改稿的第 6 页和第 13 页。

意见 5：文中第六页，“C-SPRT 理论介绍中，在法向上构造邻域，”法向是指？是方向？法向量？语意不清。

回应：很抱歉这里给您造成困惑。法向是指法向量的方向，为使表达更清楚，我们已将“法向”更改为“法向量方向”。详见修改稿的第 6 页。

意见 6：文中第十二页，((2)选题策略)“ Σ 为 θ 先验方差的协方差矩阵”，这样设计是否合理？

¹ Nydick, S. (2013). *Multidimensional mastery testing with CAT* (Unpublished doctoral dissertation). University of Minnesota.

首先， Σ 出现多次，代表不同的协方差，表述是不清晰的。此外，文中已说明 θ 的估计是极大似然估计，为何又引入先验分布？极大似然估计的标准误差与先验无关。如果是贝叶斯估计，先验的选择是非常重要的，不能简单一提，需要给出先验分布选择的理由。

回应：非常同意您的观点，也很抱歉这里存在问题。首先，针对 Σ 在文中多次出现的问题，我们进行如下修改：在公式（17）处将 Σ 修改为 Σ_k ；在公式（18）处，原文使用的 $\Sigma_{j'}$ 可以与 Σ 进行区分，因此仍保留 $\Sigma_{j'}$ 的写法；在 3.1 的被试生成中，保留 Σ 的写法；在 3.2 的选题策略中，删除 Σ 。详见修改稿的第 9 页、第 11 页和第 13 页。

另外，在选题策略中，为使得其与能力估计方法保持一致，我们将其修改为不含先验分布的原始的 D 最优选题策略，调整后的公式（26）不再包含 Σ 。相应地，我们重新进行了模拟研究，并更新了相关结果。详见修改稿的第 13 页。

意见 7：文中第九页，新规则将被试能力的一系列估计值作为一个“类”，这个“类”具体指什么，请详细描述。

回应：非常感谢您的建议。这个“类”指的是对某一名被试的多个能力估计值的集合，它在很大程度上能够勾勒出被试真实能力值在能力空间中所处的范围。

本文基于聚类分析的思想，提出基于马氏距离的 Mahalanobis-SPRT 规则。在测验初期，由于被试作答的题目数量较少，对被试能力的估计往往不够准确。但是，如果将多个能力估计值看作一个集合（或称作一个“类”），就可以大致描绘出真实能力值所处的范围。此时，因为单个的能力估计值不准，而这个“类”所确定的范围相对准确，所以使用分界曲线上到该集合的马氏距离最近的点作为 $\hat{\theta}_0$ （即 Mahalanobis-SPRT 方法的做法）就比 P-SPRT 中直接使用分界曲线上到 $\hat{\theta}_i$ 的欧氏距离最近的点要更合理。下图是某名被试在测验过程中，对被试能力的估计值随被试作答题目数量变化的展示。其中，绿色的点代表该被试能力的真值，蓝色越深表示得到该能力估计值时被试作答的题目数量越多。

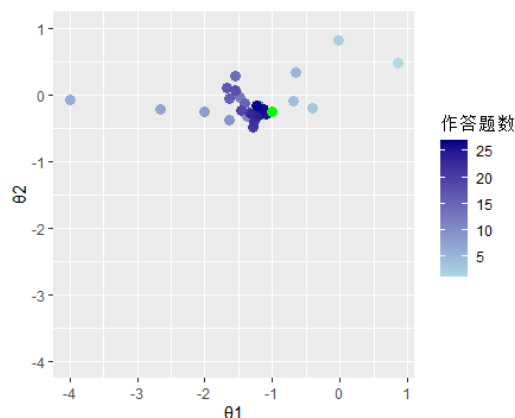


图 2 被试能力估计值随作答题数的变化图

可以看到，随着测验的进行（即被试作答的题目数量的增加），被试能力的估计值逐渐靠近真值。而在测验初期，即作答题数较少时，被试能力的估计值并不准确，但是多个估计值所划定的范围是相对准确的。我们已经将上述观点或内容补充到文章的修改稿中，详见修改稿的第 9 页。

意见 8：文中第十二页，公式(25)的参数“ \mathbf{a} ”应改为 \mathbf{a}_j 。

回应：非常抱歉这里存在错误。我们已将公式（25）中的“ \mathbf{a} ”修改为“ \mathbf{a}_j ”。详见修改稿的第 13 页。

.....

审稿人 2 意见：

本文稿本研究在对多维计算机化分类测验的终止规则进行系统梳理的基础上，提出两种新规则。问题具有研究意义，研究方法恰当。具体的修改意见如下。

回应：非常感谢您对本研究在选题意义、研究方法方面给予的积极评价。对于您提出的修改意见，我们将在下面逐条回复。

意见 1：摘要研究方法和研究结果介绍不够详细。

回应：非常认同您的建议，我们在摘要部分丰富了对研究方法及结果的描述。详见修改稿的第 3 页。

意见 2：引言第一句话，表达是否准确？CCT 是一种新兴测验形式？它应该是一种有特殊

的 CAT，与一般 CAT 相比，只是评价结果的方式不同而已；另外，评价结果为掌握和未掌握，换为达标与未达标似乎更为准确。

回应：非常感谢您的建议。如您所言，CCT 的确可以被理解为一种特殊的 CAT。为使描述更准确，我们对原文进行了修改，详见修改稿的第 3 页。此外，我们非常认同您对于评价结果的意见。在修改稿中我们统一将分类结果中涉及的掌握与未掌握替换为达标与未达标。

意见 3：引言第二句“使用因人而异的选题策略和终止规则”表达是否准确，选题策略和终止规则应该是一样的，只是在相同的规则下针对不同被试产生的结果不用而已。

回应：非常感谢您的建议。为使描述更加准确，我们将这句话改为“使用自适应的选题策略和终止规则”。详见修改稿的第 3 页。

意见 4：引言第三句，前半段说的各种测量模型，后半段又说的项目反应理论，表达上是不是应该一致？

回应：非常认同您的观点。为使表述保持一致，并增加准确性，我们将“测量模型”修改为“心理测量理论”。详见修改稿的第 3 页。

意见 5：引言第二段第一句“或”是不是应该改用为“和”？

回应：非常感谢您的建议。我们已将“或”改为“和”。详见修改稿的第 3 页。

意见 6：引言第二段第二句“由此，终止规则是区分两者的一项主要特征”，表达是否准确？应该是能力估计或判定方式才是二者的主要区别吧。CAT 中也可以有变长度的，而且 CCT 的终止规则似乎也可以用在 CAT 中。

回应：感谢您的建议，原文中的描述确实不够严谨。根本上讲，CCT 与 CAT 的主要区别在于测验目的：CAT 的目的是得到对被试能力的准确估计结果，而 CCT 只要求输出对被试的等级划分。正是由于测验目的上的不同，CCT 与 CAT 对各自的终止规则提出了不同的要求。具体地说，CCT 的终止规则在本质上是要完成一个对被试的分类问题，但 CAT 终止规则的核心任务则是控制能力估计误差。因此，CCT 与 CAT 的终止规则并不相同。此外，判定方式应当是包含于终止规则的。也就是说，CCT 的终止规则既包含假设检验及对应统计量的

构造，也包含最终对被试等级进行的判定。我们已将上述部分内容补充到修改稿中，详见修改稿的第 3 页。希望您对我们的回应和修改满意。

意见 7：本文讨论的终止规则和已有的不定长终止规则是针对二分类还是多分类的情况，应该在引言部分或者 2.2.1 开头讲清楚。

回应：非常感谢您的建议。我们在原文引言部分第二段第 4 行中已提到过“两分类测验的背景下”。为使相关描述更明显，我们在引言第二段最后增加了“二分类测验情境下”的相关描述。详见修改稿的第 4 页。

意见 8：引言第三段“迄今，关于 MCCT 的研究较少，只有少数研究者将特定的似然比规则从单维情境推广至多维情境”，有少数研究推广到多维情景，因此，应该在句末加上相关研究的文献作证。

回应：感谢您的建议。我们已增添相关研究文献。详见修改稿的第 4 页。

意见 9：多维情况下到底是能力分界曲线还是曲面？文稿多次提到，如果两个都可能，那么什么情况下是曲线，什么情况下是曲面，应该说清楚。

回应：非常感谢您的建议。多维情况是指维度大于等于 2 的情况。在二维空间中，使用一条曲线即可将空间分割为两个互斥的部分，因此二维情境与能力分界曲线相对应；而在三维及以下的空间中，就需要使用一个曲面才能将空间分割为两个互斥的部分，此时就对应能力分界面。我们在其第一次出现的地方增加了相关说明。详见修改稿的第 4 页。

意见 10：公式（2）上面一句“被试对 j' 道”建议用大写 J 表示作答的 J 个项目。

回应：感谢您的建议。由于在原文公式（10）下方的段落中， J 已经被赋予“最大测验长度”的含义，因此此处选择使用 j' 以便于区分。希望您理解并支持我们的决定。

意见 11：2.2.1 部分第一段话，（2）确定不同类别（容易产生混淆以为是能力类别，实质上是不同的等级类别吧，所以考虑改成“等级”怎么样？）间的能力阈值。

回应：非常感谢您的建议。我们已经将文中可能引起歧义的“类别”改为“等级”，详见修

改稿的第 5 页。

意见 12: 2.2.1 部分第一段话,“(3) 在处给定一个临域,即”中公式有误。

回应: 十分抱歉这里存在错误。该公式已改为 “ $(\theta_0 - \delta, \theta_0 + \delta) \equiv (\theta_l, \theta_u)$ ”, 详见修改稿的第 5 页。

意见 13: 2.2.1 部分第一段话,第(4)中的分类函数是什么样的,可否举个例子在此?

回应: 感谢您的建议,我们已在文中增加相应的例子。修改后为“其中 $g(\theta)$ 为分类函数,具体可分为补偿性的分界曲线,如 $g(\theta) = \theta_1 + \theta_2 = 0$;非补偿的分界曲线,如 $g(\theta) = \begin{cases} \theta_1 = 0, \theta_2 \geq 0, \\ \theta_2 = 0, \theta_1 > 0 \end{cases}$ ”。详见修改稿的第 5 页和第 6 页。

意见 14: 公式(4)上面一段话中“C-SPRT 的基本思路是使用约束在分界曲线上的能力估计值替代能力分界点,”与“C-SPRT 算法首先将在边界函数上计算得到的能力参数极大似然估计值作为阈值的估计,”的表达应该是对应的吧?读者不是很明白。

回应: 非常同意您的观点,这两句表达是对应的。为使两处描述更统一,便于读者理解,我们将第二句话修改为“C-SPRT 方法首先把在分界曲线上计算得到的能力参数的估计值作为能力分界点的估计”,详见修改稿的第 6 页。

意见 15: 公式(4)上面一段话第一句是“法向量”少了“量”字?

回应: 很抱歉这里给您造成困惑。法向是指法向量的方向,为使表达更清楚,我们已将“法向”更改为“法向量的方向”。详见修改稿的第 6 页。

意见 16: 公式(4)上面一段话中即,分子分母如何求?

回应: 分子部分 $\nabla g(\hat{\theta}_0)$ 中, ∇ 为哈密顿算子,即 $\nabla \equiv \frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k}$,其中 $\mathbf{i}, \mathbf{j}, \mathbf{k}$ 分别为沿 x 轴, y 轴和 z 轴正向的单位向量。因此, $\nabla g(\hat{\theta}_0)$ 就表示一个在点 $\hat{\theta}_0$ 处与分界曲线垂直的向量。分母 $\|\nabla g(\hat{\theta}_0)\|$ 表示向量 $\nabla g(\hat{\theta}_0)$ 的模长,即该向量中各坐标值的平方和的算术平方根。由此, $\frac{\nabla g(\hat{\theta}_0)}{\|\nabla g(\hat{\theta}_0)\|}$ 即表示一个在点 $\hat{\theta}_0$ 处与分界曲线垂直的单位向量。需要说明的是,本文中引入 ∇ 并使

用 $\frac{\nabla g(\hat{\theta}_0)}{\|\nabla g(\hat{\theta}_0)\|}$ 只是为了使记号更简洁，关于其含义的理解可以参看原文中的文字描述。

意见 17：公式（10）下面提到“Wald-Wolfowitz 定理表明”，请问定理的具体内容是那一部分？

回应：很抱歉给您造成困惑。原文中“Wald-Wolfowitz 定理表明”后的部分即为定理内容。具体地说，Wald-Wolfowitz 定理原文是：在观测可以持续直至满足终止规则的情况下，SPRT 是具有同等检验力的检验中所需观测个数最少的假设检验，即最优序贯检验。在本文的背景下，“观测可以持续”即是“测验可以持续”。

意见 18：公式（10）下面提到“因此在传统 CCT 中，一般通过事先设定最大测验长度以满足上述现实需要。于是，研究者在设计 MCCT 终止规则时也直接沿用这一方法，若达到最大长度 J 时测验仍未结束。”此句表达不准确，应该是所有不定长测验都有一个附加结束的强制条件。

回应：非常感谢您的建议。为使表达更准确，我们将原文修改为“于是，研究者在设计不定长的 MCCT 终止规则时也沿用这一附加的强制结束条件。这就是说，若达到最大长度 J 时测验仍未结束”。详见修改稿的第 7 页。

意见 19：公式（4）、（12）是否有误，比如（12）中应该是与分界点最接近的，还是最小的绝对值差值？公式（17）下面一段讲到“那么他/她在作答完道题目后可得到一组能力估计值 $\hat{\theta}_{ij}$ ($j = 1, 2, \dots, j'$)”。假设这组能力估计值的均值为 $\mu_{j'}$ 、协方差为 $\Sigma_{j'}$ ”。是指每个项目作答完后，各个维度的能力均值向量，还是所有维度能力在每个项目作答后的均值向量，是哪些向量间的方差协方差？

回应：非常感谢您的建议。公式（4）表示：把 Θ_0 上，使得 $\log L(\theta | Y_{ij'})$ 取最大值的点记为 $\hat{\theta}_0$ ；公式（12）表示：把 Θ_0 上，与 $\hat{\theta}_i$ 距离最近的点记为 $\hat{\theta}_0$ ，也就是将 $\hat{\theta}_i$ 投影至 Θ_0 上，并将投影得到的点记为 $\hat{\theta}_0$ 。其中， $\|\cdot\|$ 表示欧几里得范数，衡量的是欧氏空间中的距离。为避免引起歧义，我们将 $\|\cdot\|$ 修改为更完整的记号（即 $\|\cdot\|_2$ ），并增加了对 $\|\cdot\|_2$ 的解释。

在公式（17）下面一段的相关描述里，得到被试 i 的一组 p 维能力估计值为 $\hat{\theta}_{ij} = (\hat{\theta}_{ij1}, \hat{\theta}_{ij2}, \dots, \hat{\theta}_{ijp})$ ($j = 1, 2, \dots, j'$)。由此，这一组 p 维的能力估计值的均值为 $\mu_{j'} =$

$(\sum_{j=1}^{j'} \hat{\theta}_{ij1}/j', \sum_{j=1}^{j'} \hat{\theta}_{ij2}/j', \dots, \sum_{j=1}^{j'} \hat{\theta}_{ijp}/j')$, 也就是被试 i 在作答 j' 道题目后在 p 个维度各自的能力均值构成的一个 p 维向量。方差协方差矩阵是这组能力估计值, 即 $\hat{\theta}_{ij}$ ($j = 1, 2, \dots, j'$) 间的协方差矩阵。

为帮助读者理解, 我们已经将上述的一些内容补充到修改稿中。详见修改稿的第 6 页、第 7 页和第 9 页。

意见 20: 则公式 (18) 到底是每个项目后的能力估计值向量, 还是作答完 j' 项目后的能力估计值向量? 多维情况下能力向量本来就是多维向量, 这个地方是如何解释的, (18) 中代表多个多维向量?

回应: 十分抱歉这里给您造成困惑。在公式 (18) 中, 使用的是作答完 j' 道题目后的能力估计值的均值向量。如在意见 19 的回应中所述, 这个均值向量是一个 p 维向量, 因此并不存在多个多维向量。为避免歧义, 我们已经对公式 (18) 前面的描述进行修改, 详见修改稿的第 9 页。

意见 21: 公式 (18) 是不是有误? Min 的结果是其中的能力点, 还是最短的马氏距离?

回应: 很抱歉给您造成困惑。argmin 是对函数求参数的函数, 具体到公式 (18) 中, argmin 的结果是 Θ_0 中, 使得所求的马氏距离最短的能力点。

意见 22: 研究在 6 种实验情景下考察各个终止规则的 PCC 和 ATL, 但是结果是有的终止规则 ATL 短, PCC 相对低; 有的 ATL 长, PCC 相对较高? 很难实现同等条件下的比较。这个终止规则更像是能力分类的标准, 有没有考虑传统 CAT 中不定长终止规则和此处的终止规则结合适用。换句话说, 要么在测量精度达到一定标准, 要么满足本文的终止规则时结束测验, 是不是可以将所有规则的精度定在相同水平上, 然后比较 ATL 的优劣?

回应: 非常感谢您的建议。在包括本文在内的大多数关于 CCT 的研究中, PCC 一般指的是被正确分类的被试占有所有被试的比例。以本研究为例, PCC 是指在每种实验条件下, 模拟的 3000 名被试中被正确分类的被试所占的比例。也就是说, PCC 所代表的精度是针对所有被试而言的, 即整个 MCCT 系统的精度。但是在不定长 CAT 的终止规则中, “测量精度达到一定标准” 的规则的含义一般是指对当前被试的能力估计的误差在一定范围之内时, 即可

结束测验。这是针对每一个被试而言的，它衡量的是对被试能力的参数估计的精度。在 CAT 中，研究者之所以使用这一终止规则，本质上还是由 CAT 的目的（也就是要尽可能准确的测量被试能力）所决定的。但是对于 CCT 而言，由于其目的在于分类，并不需要刻意限制被试能力的参数估计的精度。因此，CAT 的终止规则很难结合到 CCT 中。

此外，的确如您所说，在进行不同终止规则的比较时，各个方法所得到的 PCC 与 ATL 均不一致，很难进行一个统一的比较（除非某种方法，比如新提出的 M-SCGLR，同时有更高的 PCC 和更小的 ATL，此时其一定更优）。为此，本研究也引入了一个综合的评价指标，即平均损失。这一指标综合考虑了 ATL 及 PCC，具体地说，PCC 越大，ATL 越小，平均损失就越小；PCC 越小，ATL 越大，平均损失就越大。为强调这一点，我们对损失的相关描述进行了补充，并更多地使用平均损失（即图 2 的结果）用以比较 6 种终止规则的表现。详见修改稿的第 14 至第 17 页。希望您能理解并满意我们的决定。

意见 23：研究一，有没有考虑将固定长度的分类测验加上作为基准进行比较？也就是说将定长测验长度设置本文的 ATL 对应的几种情况，看看定长的结果和不定长的结果相比，是否更具优势？

回应：非常感谢您的建议。首先，Kingsbury 和 Houser（1993）²在 CAT 中的研究表明，变长的 CAT 在测验效率、精度以及能力估计的收敛情况上均优于定长 CAT。尽管 CCT 与 CAT 在终止规则上有明显不同，但由于其在能力估计的过程上基本一致，因此可以预计定长 CCT 的效果不会比变长 CCT 的结果更有优势。其次，对变长 CCT 而言，使用不同终止规则，在不同的测验条件下，其 ATL 的值基本不会相同。如果需要使用定长 CCT 与之比较，则很难找到一个统一的基线，而需要在每个测验条件下对每种方法都构造一个与其 ATL 相等的定长 CCT。最后，在每个测验条件下对每种变长终止规则构造与其 ATL 完全相同的定长 CCT 在实际上是不可能的，因为 ATL 是平均测验长度，因此其值在大多情况下并非整数，此时就无法找到与之完全对应的定长 CCT。希望您能够理解我们的决定。

意见 24：3.3 部分第二段“Finkelman（2010）定义的 loss 是对某次测验的测验精度和效率的综合评价指标”请说清楚 R 的范围？Loss 函数中的值的含义，是在相同测验条件下比较该函数的值，还是不同测验条件下都可以比较？

² Kingsbury, G. G., & Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice*, 12, 21-27.

回应：非常感谢您的建议。由于 R 所代表的是“将某一名参与测验的被试错误分类”的惩罚，因此，通常意义上 R 的取值范围应大于等于 0。而且 R 的值越大表示该测验对错误分类的惩罚越大，即对精度的要求越高。根据公式（27）， $\text{loss} = R \times \mathbf{1}_W + K$ 。其中， $\mathbf{1}_W$ 表示错误分类的示性函数，在对被试错误分类时取值为 1，在没有错误分类时取值为 0。在某次测验中，如果将被试错误分类，则 loss 的值为惩罚 R 的值与该次测验的长度之和；如果将被试正确分类，则 loss 的值就等于该次测验的长度。相关内容我们已经补充到修改稿内，详见修改稿的第 13 页。

此外，在固定 R 之后，损失只与 PCC 和 ATL 有关，因此在不同测验条件下也可以对其进行直接比较。但需要说明的是，在正文的图 2 中所展示的是标准化后的平均损失（这是为了更加清晰展示各个终止规则的平均损失的相对关系随 R 变化的趋势）。这一标准化后的平均损失反映的是同一测验条件下，不同终止规则的平均损失的相对大小关系，只能在相同测验条件下进行比较。详见修改稿的第 15 页和第 16 页。

意见 25：3.3 部分最后一句话“在公式（27）中固定 R 后， P 的值与 R 相等。”“ P 的意思就是 R 吗？”

回应：非常感谢您的建议。由于在公式（27）中固定 R 后，再对其进行平均才能得到公式（28）。因此原文中公式（28）的 P 就等于 R 。为使描述更清晰，我们已经将 P 统一修改为 R 。详见修改稿第 13 页、第 14 页和第 16 页。

意见 26：第 4 部分第二段第一句“总体而言，各终止规则的 PCC 与 ATL 呈现出正相关的趋势（即平均测验长度越长，测验分类精度也越高）”这个结果似乎看不出来，体现在图上面的是不同终止规则的情况，测验长度越长，精度越高，而非相同规则的规律。

回应：很抱歉这里的描述不够准确。此处已修改为“总体而言，不同终止规则的 PCC 与 ATL 呈现出正相关的趋势（即平均测验长度越长，测验分类精度也越高）”。详见修改稿的第 14 页。

意见 27：自检报告种第一条“（2）首次将随机缩减的广义似然比规则（SCGLR）推广到多维情境中，并提出 M-SCGLR。模拟结果显示：该方法在不影响分类精度的前提下能大幅缩减测验长度，尤其是在题目间多维的题库条件下，该方法在分类精度和平均测验长度等指标上全面优于已有的多维随机缩减方法（M-SCSPRT）。”是如何体现的？读者未能从结

果中总结这一条结论。

回应：非常感谢您的建议。我们将对这段话中的结论进行逐句解释：

（1）“该方法在不影响分类精度的前提下能大幅缩减测验长度”是相对于未使用随机缩减技术的 4 种方法（C-SPRT、P-SPRT、Mahalanobis-SPRT 及 M-GLR）而言的。采用随机缩减技术的 M-SCGLR 的 ATL 在各条件下均明显小于上述 4 种方法。同时，M-SCGLR 方法在 PCC 上虽然较上述 4 种方法有一定下降，但这种下降是由于使用随机缩减技术而导致的合理下降（具体表现为其下降幅度远小于 M-SCSPRT 方法）。

（2）“该方法在分类精度和平均测验长度等指标上全面优于已有的多维随机缩减方法（M-SCSPRT）”。这个观点在新的模拟结果中不再成立。根据新的实验设计所得到的结果显示，M-SCGLR 方法的分类精度较 M-SCSPRT 有很大提高（在非补偿分界曲线的情境下甚至接近未使用随机缩减技术的规则），而平均测验长度增加的并不多。因此，我们对这句话也进行了相应的修改。

意见 28：CCT 中终止规则实质上就是分类准则，是吗？当能得到明确的分类结果时就可以结束测验了。也可以是固定测验长度，测验结束时运用分类准则进行判定分类结果？因此，本文需要说明分类规则和终止规则的差别。

回应：由于 CCT 终止规则的最终目的是对被试进行分类，因此从本质上说，CCT 中的终止规则可以被理解为分类准则。但需要说明的是，如 2.2.1 第一段所介绍的，一个 CCT 的终止规则不仅包含最终对被试分类的规则，而且也包含最初的假设检验、无差别区间及检验统计量的构造。

此外，的确可以固定测验长度，在被试作答指定长度的题目后就结束测验，并给出对被试的分类（类似于本文中达到最大测验长度时的规则）。但是固定长度的测验并不完全符合“自适应”的思想，而且在实践中会导致部分早就可以完成分类的被试被迫多作答一定题目，或者一些还不能进行准确分类的被试被迫停止测验。希望您对我们的回应感到满意。

第二轮

审稿人 1 意见：

本篇稿件尚存在以下问题：

意见 1: 文中 2.2.1 的(3), 在介绍 M-GLR 方法时强调了不需要确定 $\hat{\theta}_l$ 和 $\hat{\theta}_u$, 那在判别准则中出现的 $\hat{\theta}_l$ 和 $\hat{\theta}_u$ 又是如何计算的, 前后矛盾, 请进行修改。而且此广义似然比方法(公式(13)介绍)适用于固定的样本量, 对于样本量不确定的情况, 此方法应该是不合适的。请认真验证理论的合理性, 并附上相应的过程或文献。

回应: 非常感谢您的宝贵建议。首先, 在 M-GLR 方法中, 的确不需要确定 $\hat{\theta}_l$ 和 $\hat{\theta}_u$ 。正文所述的“P-SPRT 和 M-GLR 的判断准则都与 C-SPRT 的判断准则(即公式(11))一致”并不意味着 M-GLR 方法中包含 $\hat{\theta}_l$ 和 $\hat{\theta}_u$ 。M-GLR 方法是根据公式(13)构造广义似然比统计量 $C_{ij'}$, 这一计算过程并不涉及 $\hat{\theta}_l$ 和 $\hat{\theta}_u$ 。C-SPRT 和 P-SPRT 则是依据公式(7)构造序贯似然比统计量 $C_{ij'}$, 其中包含 $\hat{\theta}_l$ 和 $\hat{\theta}_u$ 。在构造完统计量 $C_{ij'}$ 之后, 三种方法都是通过与 C_l 、 C_u 或 C_0 进行比较得到测验结果, 即按照公式(11)所定义的规则对被试做出分类判断。很抱歉这里给您造成误解, 我们对原文进行了进一步修改, 详见修改稿正文的第 8 页。

其次, 我们认真查阅相关文献后发现广义似然比方法也可以用于样本量不确定的情况。如 Bartroff, Finkelman 和 Lai (2008)³首次将 GLR 方法引入变长的 CCT 中, 并对此情境下 GLR 的良好性质进行详细证明。之后, Nydick (2013)⁴将这一方法推广至变长的 MCCT, 得到本文公式(13)所示的 M-GLR。与单维 GLR 方法相比, M-GLR 只是将 $C_{ij'}$ 中求极值的集合由单维的区间变为多维空间中的区域, 其性质并没有变化。此外, Huebner 和 Fina (2015)⁵以及 Thompson (2011)⁶等人也都在变长 CCT 的情境下对 GLR 方法进行过探讨。非常感谢您指出这一点, 我们在引言部分首次介绍 GLR 方法时以及在“2.2.1 (3)”介绍 M-GLR 时均补充了对这一内容的描述, 详见修改稿正文的第 4 页和第 8 页。

意见 2: 文中 2.3.1, “记被试 i 作答完第 j 道题目后得到的能力估计值为 $\hat{\theta}_{ij}$, 那么他/她在作答完 j' 道题目后可得到一组 p 维的能力估计值 $\hat{\theta}_{ij} = (\hat{\theta}_{ij1}, \hat{\theta}_{ij2}, \dots, \hat{\theta}_{ijp})$ ($j = 1, 2, \dots, j'$)。”此处如此计算的协方差 $\Sigma_{j'}$, 并不是不同维度能力之间的相关系数。该相关矩阵只是一个被试作答不同题目后能力估计值的相关性。不同能力之间的相关性应是被试总体层面的, 被试个体各个能力不存在关系。读者觉得第一次审稿意见提出的这个问题并没有得到解决, 请认真

³ Bartroff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika*, 73, 473-486.

⁴ Nydick, S. (2013). *Multidimensional mastery testing with CAT* (Unpublished doctoral dissertation). University of Minnesota.

⁵ Huebner, A. R., & Fina, A. D. (2015). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior Research Methods*, 47, 549-561.

⁶ Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, & Evaluation*, 16, 1-7.

阅读并修改。

回应：非常感谢您的建议。我们非常认同您所说的“该相关矩阵只是一个被试作答不同题目后能力估计值的相关性”。这个协方差只是在计算马氏距离的过程中需要使用到，对于不同的被试及不同的测验阶段，该协方差的值是变化的，自然不可能代表总体层面上的不同能力之间的相关。事实上，在修改稿正文中，我们也将该协方差定义为“这一组 p 维能力估计值的协方差”，并没有再赋予其“不同维度之间的相关”这一含义。

为使描述更加准确，我们对相关表述进行了修改并在 $\mu_{ij'}$ 和 $\Sigma_{ij'}$ 中增加了下标 i 。修改后的句子为“记被试 i 作答完第 j 道题目后得到的能力估计值为 $\hat{\theta}_{ij}$ ，那么他/她在作答完 j' 道题目后可得到 j' 个 p 维的能力估计值 $\hat{\theta}_{ij} = (\hat{\theta}_{ij1}, \hat{\theta}_{ij2}, \dots, \hat{\theta}_{ijp}) (j = 1, 2, \dots, j')$ 。由此，这名被试的 j' 个能力估计值的均值为 $\mu_{ij'} = (\sum_{j=1}^{j'} \hat{\theta}_{ij1}/j', \sum_{j=1}^{j'} \hat{\theta}_{ij2}/j', \dots, \sum_{j=1}^{j'} \hat{\theta}_{ijp}/j')$ 、协方差为 $\Sigma_{ij'}$ ”，详见修改稿正文的第 10 页。希望您对我们的修改满意。

意见 3：文中 2.3.1，“因为单个能力估计值不准确，而这个“类”所确定的范围相对准确，所以使用分界曲线或曲面上到该集合的马氏距离最近的点作为 $\hat{\theta}_0$ （这也正是 Mahalanobis-SPRT 方法的做法）就比 P-SPRT 中直接使用分界曲线或曲面上到 $\hat{\theta}_i$ 的欧式距离最近的点要更合理。”这句话是如何得出的？根据这个“类”所确定的能力估计值得范围并不准确。被试作答的题目数越多，对被试的能力估计值越准确，和上面那句话不是一种情况，不能作为解释的依据。

回应：非常抱歉这里没有描述清楚，让您造成误解。首先，在测验的初期阶段，由于被试作答的题目较少，得到的被试能力估计值往往不准确。但是在 Mahalanobis-SPRT 方法中，我们是将这些并不准确的多个能力估计值看作一个集合，并将分界曲线或曲面上到该集合的马氏距离最近的点作为 $\hat{\theta}_0$ 。需要指出的是，到这个集合的马氏距离实际上也是到这个集合的中点（即这组能力估计值的均值）的马氏距离。‘这个“类”所确定的范围相对准确’的含义是：测验初期的能力估计值是在真值附近上下波动，而并非一致地高于或低于真值，所以这个“类”的中心，也即多个能力估计值的均值，往往就更加接近真值。为帮助读者理解，我们增加了修改稿正文中的图 1 并补充了相关描述，详见修改稿正文的第 9 页和第 10 页。

意见 4：关于公式(21)、公式(22)和公式(23)的理论推导过程，请附在附录中。

回应：感谢您的建议。我们在附录中增加了相关推导，详见修改稿正文的第 20 页和第 21 页。

意见 5：文中 3.2 中 (3)，能否详细介绍补偿性分界曲线和非补偿性分界曲线有什么区别和联系，各自的定义是什么样的。

回应：非常感谢您的建议。我们在“3.2 的 (3)”中增加了对补偿性分界曲线和非补偿性分界曲线的介绍，详见修改稿正文的第 14 页。

意见 6：文中 4 结果与结论中，“对于本文提出的 Mahalanobis-SPRT，在补偿性分界曲线的情境下，其总体上具有较高的 PCC。”未能从图中得到相应的结论。

回应：非常感谢您的建议。在补偿性分界曲线的情境下，在题目间多维时，Mahalanobis-SPRT 方法的 PCC 仅仅略低于表现最好的 P-SPRT 方法（表现为在正文图 2 第一行中，Mahalanobis-SPRT 在纵坐标上仅比 P-SPRT 方法略低，而不低于其他方法）；在题目内多维时，Mahalanobis-SPRT 方法具有 6 种方法中最高的 PCC（表现为在正文图 2 第三行中，Mahalanobis-SPRT 在纵坐标上不低于其他方法）。因此，认为该方法在补偿性分界曲线的情境下，总体上具有较高的 PCC。我们补充了相关描述，详见修改稿正文的第 16 页。

.....

审稿人 2 意见：

修改稿有较大进步，行文还是存在许多表述不清楚的地方，结果呈现过于复杂，不简洁，也未能体现方法的优势。具体的意见如下。

回应：非常感谢您对修改稿的积极评价。对于您提出的修改意见，我们将在下面逐条回复。

意见 1：摘要的结果部分，两个新方法的特点太过于局限，这是分类测验，而且研究的是终止规则，那么测验的效率才是研究的核心。

回应：非常感谢您的建议。我们在摘要的结果部分增加了两种新方法在测验长度上的表现，详见修改稿正文的第 3 页。

意见 2：引言第一段末尾的引用是 2015 年的情况，那么近五年来，研究问题和方向是否发生了变化？引言第二段“最早的似然比，...，随机缩减方法（Stochastically Curtailed GLR, SCGLR）”这一部分的写作思路是不是可以将似然比的方法分为序贯似然比和广义似然比，及其基于随机的扩展。另外，SPRT 和 GLR 之间的差异没有叙述。

回应：十分感谢您的建议。我们首先在引言的相应位置补充了近 5 年关于 CCT 的新文献，对它们进行阅读后发现：相关研究都选择项目反应理论作为理论基础，而且其中绝大多数研究探讨的都是可变长度的 CCT。此外，我们非常认同您对引言第二段部分内容的写作思路，并按照您所提出的思路对相应内容进行了重新组织，详见修改稿正文的第 3 页和第 4 页。

意见 3：引言部分，只是介绍各种方法的思想，缺乏对各种方法的优势和不足的评论，也没有说清楚本研究使用方法的源起，是基于哪些不足之处建构起来的。还有，有研究比较了随机缩减和没有随机缩减的两类方法之间的表现吗？

回应：非常感谢您指出这一点。根据您的建议，我们在引言部分增加了对相关方法的优势和不足的描述。此外，之前有许多研究在单维情境下比较了使用随机缩减的方法与未使用随机缩减的方法的表现，大多数研究的结果表明：使用随机缩减的方法能够提高测验效率。我们在修改稿正文的第 4 页补充了这些文献内容。

意见 4：2.2.1 中第一段第三行临域应该是邻域。

回应：十分抱歉这里存在错误。我们已将“临域”修改为“邻域”，详见修改稿正文的第 6 页。

意见 5：2.2.1 中第一段“需要定义能力分界曲线甚至曲面才能将不同类别的被试区分开来。由此，单维情境下的能力分界点 θ_0 就变为多维空间中的能力分界曲线或曲面 $g(\theta) = 0$ ”是不是应该将“甚至”改为“或”，在“分界曲线或曲面 $g(\theta)=0$ ”后面加上“的解”。

回应：非常感谢您的建议，我们已在正文中将“甚至”改为“或”，详见修改稿正文的第 6 页。此外， $g(\theta) = 0$ 代表的是一个曲线或曲面，比如 $g(\theta) = \theta_1 + \theta_2 = 0$ 表示的就是 $\theta_1 + \theta_2 = 0$ 这样一条直线，并不涉及“解”的概念。

意见 6: 全文应该是“欧氏”不是“欧式”。

回应: 十分抱歉存在这样的错误, 我们已经重新检查全文, 并将文中表述统一为“欧氏”(详见修改稿正文的第 7、第 8 页以及第 9 页)。再次感谢您的建议。

意见 7: “ $\theta_\delta = \frac{\nabla g(\theta_0)}{\|\nabla g(\theta_0)\|_2}$, 其中 $\|\cdot\|_2$ 表示欧几里得范数”中分子的计算不清楚, 应该加以解释。

回应: 非常感谢您的建议。我们在文中增加了对分子部分的解释说明, “其中 ∇ 为哈密顿算子, 表示微分运算”, 详见修改稿正文的第 6 页。需要说明的是, 本文无意增加读者的认知负荷, 文中引用该算子对 θ_δ 进行展开只是为使记号简洁, 关于其含义的理解可参见正文描述。

意见 8: 2.2.1 的最后一部分“但都是基于公式 (3) 所建立的假设检验, 而且仅依赖构造的似然比统计量进行判断, 因此 P-SPRT 和 M-GLR 的判断准则都与 C-SPRT 的判断准则(即公式 (11))一致。”表述不准确。首先, (3) 是假设检验提出的假设, 其次 P-SPRT 并不是依据似然比来构建的。

回应: 感谢您指出这一点, 我们已对该部分的表述进行修改, 以求表达更加严谨, 详见修改稿正文的第 8 页。另外, P-SPRT 方法实际上也是基于似然比统计量构建, 如正文第 8 页第一段所述, “P-SPRT 与 C-SPRT 方法的唯一区别在于它采用不同方法将分界曲线转换为分界点”。

意见 9: “预分类的判断准则与最大测验长度下似然比检验的判断准则一致, 即

$$\begin{cases} D_{j'} = n, & C_{ij'} \leq C_0 \\ D_{j'} = m, & C_{ij'} > C_0 \end{cases}$$

部分文字不能反应公式的内容, 二者不一致。

回应: 感谢您的建议, 这两部分内容实际上是一致的。为了帮助阅读、避免歧义, 我们在修改稿正文的第 7 页对“最大测验长度下似然比检验的判断准则”进行了明确。

意见 10: 2.3.1 部分“在测验初期, 尽管单个被试能力估计值并不准确, 但是如果将多个能力估计值看作一个集合(或称作一个“类”), 就可以大致描绘出被试真实能力值所处的范围。”本文的做法实质上就是多维的马氏距离算法呀, 并没有用到什么多个能力估计值看作

一个集合。因为单维的情况下，也是算能力估计值和均值的距离吧。

回应：非常感谢您的建议。对于本文所提出的 Mahalanobis-SPRT 方法，正如其名，该方法的确是基于马氏距离的思想而提出。但与此同时，本研究中并没有选择分界曲线上与被试的单个能力估计值的马氏距离最近的点作为 $\hat{\theta}_0$ ，而是将被试的多个能力估计值看作一个“类”，并把分界曲线上到该“类”的马氏距离最近的点作为 $\hat{\theta}_0$ 。

意见 11：公式 18 下面的一句话“ θ_δ 为 $\mu_{j'}$ 与 $\hat{\theta}_0$ 连线方向的单位向量”表示两个向量有连线？只能是向量终点的连线，这里可表述为两个向量的差向量的单位向量？

回应：非常感谢您的建议。我们将这句话修改为“ θ_δ 为 $\mu_{j'}$ 与 $\hat{\theta}_0$ 的差向量方向上的单位向量”，详见修改稿正文的第 10 页。

意见 12：公式 22，23 中均值和方差如何计算？里面明明就是一个确定的值。

回应：非常感谢您的建议。在公式（22）与（23）中，其值实际上都是变化的。对于公式（22）和（23）而言，它们均包含对第 $j' + 1$ 到第 J 道题的求和，正如文中所述，这部分题目的选取是基于对被试能力的估计值。随着测验进行，被试能力的估计值会不断更新，导致题目选取的改变，进而导致两个公式所包含的求和项的变化。此外，公式（22）中还包含一个涉及已作答的 j' 道题目的部分，即 $C_{ij'}$ 。随着测验进行，被试作答的题目不断增加， $C_{ij'}$ 的值自然也会随之变化。为使描述更准确，我们对公式（22）与（23）进行了微调并在附录中给出了其证明过程，详见修改稿正文的第 11 页、第 20 页以及第 21 页。

意见 13：研究设计部分的建议：

回应：非常感谢您的建议。下面我们对您提出的具体意见给予回应。

意见 14：3.1 部分（1）的内容：这样生成的两个维度的区分度肯定是负相关的，但是事实上项目的多个维度之间的区分度显然不一定呈负相关，因此实验条件是否和实际吻合，是否考虑周全？为什么不直接在区分度的分布中产生两个维度的区分度？

回应：非常感谢您提出的宝贵建议。如文中所述，在本研究所采用的多维项目反应理论模型（即多维三参数逻辑斯蒂克模型）中，与单维项目反应理论模型中的题目区分度参数所对应

的是MDISC,也就是区分度参数向量($a_{j1}, a_{j2}, \dots, a_{jp}$)中各元素的平方和的算术平方根。因此,从区分度的分布中产生MDISC,再从中分别赋予两个维度的区分度可能会获得更接近真实区分度的参数值。这种做法也是参考了 Nydick (2013)⁷的研究而确定。

正如您所说,这样生成的两个维度的区分度参数之间会存在负相关的关系。我们将在今后的研究中比较不同区分度生成方式对结果的影响。相关内容已经被补充在讨论部分,详见修改稿正文的第 20 页。

意见 15: 文中用到补偿性分界曲线和非补偿性分界曲线,那么这两种分界曲线应该和使用的模型或者能力之间是否具备从补偿性来选用,而本文使用 M3PL,也用了这两类恰当吗?或者应该说明两个分界曲线的选用标准。

回应: 非常感谢您的建议。的确如您所说,当选择补偿性的 MIRT 模型时,同样为补偿性的分界曲线应该会更加与之匹配。但是,本文模拟研究的主要目的是比较各种终止规则在精度和效率上的表现。为了观察不同分界曲线对终止规则的表现可能造成的影响,本文设置了补偿性与非补偿性曲线这两个条件。关于这一点,我们在“实验部分的第 2 段”进行了说明,详见修改稿正文的第 12 页。

此外,本研究之所以选择补偿性的 M3PL 模型,是因为该模型相较于非补偿性模型的适用范围更广。事实上,在大多数关于 MCCT 及 MCAT 的研究中,所选用的也都是补偿性模型(比如,Chen & Wang, 2016⁸; Chen, Wang, Xin, & Chang, 2017⁹; Wang & Chang, 2011¹⁰)。在未来研究中,我们也可以考虑非补偿性 MIRT 模型对各种终止规则表现的影响。我们在讨论中增加了相关内容,详见修改稿正文的第 20 页。

意见 16: 研究设计中既然能力维度之间可以相关,是否会反应在题目两个区分度之间的相关?是不是区分度之间也应该具备这种关系?

回应: 非常感谢您的建议。正如对您意见 14 的回复一样,这是我们在未来研究中可以考虑

⁷ Nydick, S. (2013). *Multidimensional mastery testing with CAT* (Unpublished doctoral dissertation). University of Minnesota.

⁸ Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika*, 81, 674-701.

⁹ Chen, P., Wang, C., Xin T., & Chang, H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 70, 81-117.

¹⁰ Wang, C., & Chang, H. (2011). Item Selection in Multidimensional Computerized Adaptive Testing—Gaining Information from Different Angles. *Psychometrika*, 76, 363-384.

的内容。我们在讨论中补充了相关内容，详见修改稿正文的第 20 页。

意见 17：4 结果部分的建议：

回应：非常感谢您的建议，我们将对您下面提到的具体问题进行回应。

意见 18：第三段第二句“随着能力维度间相关系数的增加，6 种终止规则的 ATL 有减少的趋势，而 PCC 则有升高的趋势。”这个关于 ATL 的结果似乎不明显？

回应：非常感谢您的建议。正如您所说，ATL 的变化的确没有比 PCC 的变化更明显，但其确实有减少的趋势。此外，为将所有方法的结果都能够呈现在图 2 中，ATL 的坐标跨度会比较大，这也是导致“ATL 的减少趋势看起来不明显”的一个原因。为使表达更清晰，我们增加了相关的数值描述，详见修改稿正文的第 16 页。

意见 19：接下来“相比于非补偿的分界曲线，6 种终止规则在几乎所有的补偿性分界曲线情境下的 ATL 均有所下降，而 PCC 则有所升高。”原因是什么？

回应：非常感谢您的建议。这可能是因为本研究所使用 M3PL 模型是补偿性模型，与补偿性的分界曲线更契合，所以导致“在补偿性曲线情境下，各种终止规则的表现均更好”。我们已经将相关描述增加到修改稿正文的第 16 页。

意见 20：结果似乎随机缩减的方法没有不随机缩减的好，那么为什么要研究随机缩减的方法呢？

回应：非常感谢您的建议。正如您所见，如果仅按 PCC 对不同的终止规则进行衡量，随机缩减的方法的确要弱于非随机缩减的方法。但是，如果从 ATL 的角度进行衡量，大多数随机缩减的方法则要明显优于非随机缩减的方法。以本文的 2 维情境为例，非随机缩减的方法的 ATL 大多分布在 40 至 70 之间，而随机缩减的方法的 ATL 几乎全部低于 30。这也就是说，尽管随机缩减的方法损失了一定的分类精度，但是能够较大地缩短被试作答的测验长度，从而减少考生疲劳效应、练习效应的影响并节省测验成本，这对于计算机化分类测验的实际应用提供了很大帮助（Finkelman, 2008¹¹；Huebner & Fina, 2015¹²）。我们补充了相关描述，详

11 Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, 33, 442-463.

12 Huebner, A. R., & Fina, A. D. (2015). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized

见修改稿正文的第 16 页。

意见 21: R 的值为什么呈现每个人的结果, 为什么不能对被试总体求均值和方差, 呈现表格? 另外, 为什么 R 的值对前面 1000 个被试而言, 出现波动? 为什么称为标准化的平均损失? 这个数值就是每个被试的损失函数的标准分数吧。建议将图 2 改为数字表格的形式。

回应: 非常感谢您的建议。很抱歉这里给您造成误解, 本文中的 R 并没有呈现每个人的结果, 而是对被试总体的计算结果。在图 3 (修改前的图 2) 中, 横坐标表示的是 R 的取值, 而并非被试的序号。根据公式 (28) 的定义, R 的不同取值决定了平均损失中 PCC 的权重, R 越大, 分类精度对平均损失的影响就越大; R 越小, ATL 对平均损失的影响就相对更大。

意见 22: 结果部分的可读性不强, 本文提出的方法在 MCCT 下与已有方法相比没有明显的优势。

回应: 非常感谢您的建议。对于本研究提出的 Mahalanobis-SPRT 方法, 模拟研究结果显示: 在使用补偿性的分界函数时, 该方法具有较高的分类精度和与同类方法相近的测验长度。值得注意的是, 该方法能够从理论上克服已有规则在测验初期可能导致的分类结果不稳定。此外, 由于马氏距离具有“不受量纲影响”的特点, 该方法可以为今后结合反应时(response time)开展的 MCCT 终止规则研究提供一种可行路径。

对于本研究所提出的另一种终止规则, 即由 SCGLR 推广得到的 M-SCGLR 方法, 模拟研究的结果表明: 在几乎所有实验条件下, 该方法不仅在分类精度上大幅优于已有的多维随机缩减方法 (M-SCSPRT), 而且保持了较低的测验长度。

第三轮

审稿人 1 意见:

作者很好地回答了审稿人问题, 并根据审稿人所提出的关键问题进行了详细修改, 回答准确, 修改合理, 建议接收发表。

回应: 非常感谢您对本研究的积极评价。

审稿人 3 意见：

由于测验太长会引起被试的疲劳效应、练习效应等，因此在保证一定的测量精度条件下，缩短测验长度是一项有意义的研究课题，因此对于 MCCT 的终止规则的探讨有必要。文章经过两次修改以后，可能还存在以下几个问题需要澄清。

回应：非常感谢您对本研究的积极评价。对您所提出的修改建议，我们接下来一一给予回应。

意见 1：基于马氏距离的多维序贯比规则中，对于多个能力估计值的集合（记之为集合 A）与分界曲线/曲面集合（记之为集合 B）之间的距离的定义是否正确？（文章 p.10 的注解是到这个集合的马氏距离实际上也是到这个集合的中点（即这组能力估计值的均值）的马氏距离。）数学中两个集合的距离似乎是豪斯多夫距离（集合之间的距离度量），这至少可以在百度中找到。当然可以找更加专业的数学书参考。既然两个集合 A,B 之间的距离定义有问题，那么文章中相应于马氏距离的结果是不是可能也要发生变化？

回应：很抱歉给您带来误解。在原文中所要衡量的实际上是点到集合的距离，而并非两个集合之间的距离。也正是通过计算分界曲线上的各个点到“由多个能力估计值所组成的”集合的马氏距离，我们才可以选择距离最短的点作为 $\hat{\theta}_0$ 。为使表述更准确，我们对原文进行了修改，详见修改稿正文的第 10 页。

意见 2：文章中对 Fisher 信息量和信息矩阵的概念没有很好地区分，因此公式 (26) 下面的说法比较混乱，信息矩阵和信息量不可以等同对待；同样，正文 p.11 第 3 节上方关于最大化 Fisher 信息量的说法也请仔细斟酌；另外公式 (25) 表达的虽然是一个矩阵，但是这个矩阵的秩 (rank) 是一维的，因为它本质上是一个列向量和一个行向量的乘积（然后再数乘），所以它的行列式等于零。

回应：非常感谢您的建议。Fisher 信息矩阵是 Fisher 信息量在多维情境下的推广。为使表述更加准确，我们对相关内容进行了修改，详见修改稿正文的第 11 页、第 13 页和第 14 页。

此外，公式 (25) 计算的是一道题目的 Fisher 信息矩阵，如果对公式 (25) 计算行列式，其值的确为 0。但是 D 最优策略是按公式 (26) 的定义进行计算，而公式 (26) 中涉及的是 j 道题目的 Fisher 信息矩阵之和的行列式，其值并不等于 0。为使表述更准确，我们对原文进行了修改，详见修改稿正文的第 14 页。

意见 3: 对于 MIRT, 其能力维数 p 是一个重要的量, 但是模拟研究没有考虑它, 后面的讨论也没有专门提及它。一般来说, 为了达到相同的能力估计精度, p 越大, 使用的题量越大, 即测验长度越长; 所以维数 p 与随机压缩之间的关系如何, 理应考虑;

回应: 感谢您的宝贵建议。我们完全同意能力维数是一个非常重要的影响因素, 并专门在讨论部分补充了相关内容, 详见修改稿正文的第 20 页。谢谢您指出这一点。

意见 4: 文章中公式 (18) 是协方差矩阵, 但是 p.12 倒数第 3 行的矩阵是相关矩阵, p.13 表 1 使用的仍然是相关矩阵; (18) 中协方差矩阵是不是一定可逆, 何时可逆, 有必要说一声, 比如可以参考张尧庭, 方开泰的《多元统计分析》书。建议修改以后再审。

回应: 非常感谢您的宝贵建议。公式 (18) 中的协方差矩阵为 $\Sigma_{ij'}$, 是在被试 i 作答 j' 道题目后, j' 个能力估计值之间的协方差矩阵, 这和第 12 页倒数第 3 行中的 Σ 不一样。 Σ 表示的是, 被试不同维度间的能力的协方差矩阵。由于这两个维度上能力的方差均为 1, 所以第 12 页中的协方差矩阵实际上就是相关矩阵。

此外, 对于公式 (18) 中的矩阵的可逆性, 我们也在修改稿中进行了说明 (详见修改稿正文的第 10 页)。

.....

审稿人 4 意见:

本文经过两轮修改已得到了较大的改进。文章新提出了 Mahalanobis-SPRT 和 M-SCGLR 规则两种 MCCT 终止规则, 是一篇具有理论创新的论文, 新的终止规则对今后 MCCT 的实际应用具有直接的指导意义。文章基本结构清晰, 方法推导正确, 结果可信, 基本达到发表要求, 但评审人仍有几处小的修改建议:

回应: 非常感谢您对本研究的积极评价。我们将在下文中对您所提出的修改建议给予回应。

意见 1: 结果部分的撰写逻辑性仍需要提升, 作者提出了两个新的终止规则, 则应该以这两个规则为基本的重心进行结果叙述和讨论。例如, 对图 1 的描述, 大量的结果都在阐述总体的 6 个指标如何, 而读者更希望看到的是新指标相比其他指标的表现优劣与否。

回应: 非常感谢您的宝贵建议。我们在修改稿中对结果部分的结构进行重新组织, 并优先对

新规则的结果进行叙述，详见修改稿正文的第 15 页和第 16 页。

意见 2：将结果与结论分开，结论放在全文最后，做一个精简的总结即可。

回应：非常感谢您指出这一点。根据您的建议，我们在第 4 部分“结果”中只保留对结果的叙述，而将结论放在文末的第 6 部分“结论”。详见修改稿正文的第 21 页。

意见 3：结果中夹杂着讨论的内容，例如：“……这与 Thompson（2011）在单维情境下得到的结论一致”，又如：“这可能是由于本研究考虑的非补偿边界其实就是直角坐标系中构成第一象限的坐标轴，……”。类似的地方还较多，作者需注意区分结果和讨论的撰写。

回应：完全同意您的建议。我们已经将讨论的内容统一调整到第 5 部分“讨论及未来的研究方向”中，详见修改稿正文的第 19 页和第 20 页。

意见 4：文中的模拟研究结果基本上是以图的形式呈现，呈现方式方便理解，但不利于之后研究者的精确比对，作者结合文章篇幅在文中补充结果表格或以补充材料形式提供具体结果的下载链接。

回应：非常感谢您的建议。我们已经将正文的图 2 所对应的表格增加到附录 2 中，详见修改稿正文的第 22 页、第 23 页和第 24 页。

意见 5：作者提出的 Mahalanobis-SPRT 事实上并没有达到预想的效果，例如在补偿性分解曲线条件下，所有的 Mahalanobis-SPRT 的 ATL 都要高于 P-SPRT 而 PCC 却没有太大的差异。从理论上 Mahalanobis-SPRT 弥补了 P-SPRT 初期 θ 估计的问题，但结果上并没有很好的反映，正如作者所说 Mahalanobis-SPRT 可能依赖更多前期作答信息，这是否产生矛盾，建议作出更多讨论。

回应：非常感谢您的宝贵建议。本研究所设置的最大测验长度为 100，这意味着在测验结束时，往往能够得到比较准确的被试能力估计值。因此，Mahalanobis-SPRT 方法对 P-SPRT 初期的能力估计问题的弥补可能就无法很好体现。但是，当最大测验长度较小时，该方法可能会有更好表现。我们已经将相关讨论增加到修改稿中，详见修改稿正文的第 19 页。

意见 6: 图 2 中两个绿色的线条很难区分, 请作者作出调整。

回应: 非常感谢您的宝贵建议。我们对图 2 及图 3 中不同方法所对应的颜色进行了调整, 详见修改稿正文的第 15 页和第 17 页。

第四轮

审稿人 4 意见:

目前文章已得到了较好的修改, 没有更进一步的意见, 建议发表。

回应: 非常感谢您对本研究的积极评价。

.....

审稿人 3 意见:

对于“马氏距离的多维序贯比规则中, 对于多个能力估计值的集合(记之为集合 A)与分界曲线/曲面集合(记之为集合 B)之间的距离的定义是否正确?(文章 p.10 的注解是到这个集合的马氏距离实际上也是到这个集合的中点(即这组能力估计值的均值)的马氏距离。)”的回应, 审稿人还是有一点疑惑。作者以注释的形式作出解释, 但是纵使是点集到点集合的距离也不能够像作者那样定义。作为计算机化自适应测验, 当然要考虑运行速度, 所以作者采用能力估计值的平均值到分界曲线/分界曲面的距离是一种追求速度的变通做法是可以理解的(或者说作者自己就这样规定也是可以接受的), 然而作为一个集合到另外一个集合的定义恐怕不太严密。希望作者再仔细研究这个问题。

回应: 很抱歉给您带来疑惑。在修改稿中, 我们对 Mahalanobis-SPRT 规则中涉及的距离重新进行了描述。新的描述中不再涉及到集合的马氏距离, 而是点到点之间的马氏距离。

具体地说, Mahalanobis-SPRT 选择分界曲线或曲面上的点中, 到“已得到的多个能力估计值的均值”的马氏距离最近的点作为 $\hat{\theta}_0$, 即 $\hat{\theta}_0 = \operatorname{argmin}_{\theta \in \Theta_0} \|\bar{\theta}_{ij'} - \theta\|_M$ 。其中, θ 为分界曲线或曲面上的点, $\bar{\theta}_{ij'}$ 为被试 i 作答完 j' 道题目后得到的 j' 个能力估计值的均值, $\|\cdot\|_M$ 代表马氏距离。这个定义与 P-SPRT 的定义有两点不同: (1) Mahalanobis-SPRT 使用“已得到的多个能力估计值的均值”代替 P-SPRT 中的单个能力估计值; (2) Mahalanobis-SPRT 使用马氏距离作为距离的度量方式, 而非 P-SPRT 中的欧式距离。详见修改稿的第 9 页、第 10 页和第 11 页。